BIOSTATISTICS WORKSHOP:
REGRESSION
ASSOCIATION VS. PREDICTION

Sub-Saharan Africa CFAR meeting
July 18, 2016
Durban, South Africa

Regression – what is it good for?

- Explore Associations
  - Between outcomes and exposures
  - Between outcomes and exposures adjusting for confounders
  - Hypothesis Testing

- Prediction
  - Generate models to predict an outcome or event
  - Select variables to be included in prediction models
  - Generate "rules" for disease prediction

- ~~It's the solution to all our problems in medical research~~

Dementia and Memory Loss in HIV

- Explore factors that contribute to memory loss in HIV+ individuals
- Create a prediction rule for onset of dementia

- Cross-sectional Study of n=1000 HIV+ people
- Collect information on
  - **Score on memory test (continuous: higher is better)**
  - **Dementia diagnosis (binary)**
  - Age (continuous)
  - Sex (binary)
  - Clinic
  - Size of household (continuous)
  - Treatment status (categorical)
  - Years since diagnosis of HIV (continuous)

Outcomes

Predictors and covariates

Dementia and Memory Loss in HIV

- **Explore factors that contribute to memory loss in HIV+ individuals**
- Create a prediction rule for onset of dementia

- Cross-sectional Study of n=1000 HIV+ people
- Collect information on
  - **Score on memory test (continuous: higher is better)**
  - **Dementia diagnosis (binary)**
  - Age (continuous)
  - Sex (binary)
  - Clinic
  - Size of household (continuous)
  - Treatment status (categorical)
  - Years since diagnosis of HIV (continuous)

Outcomes

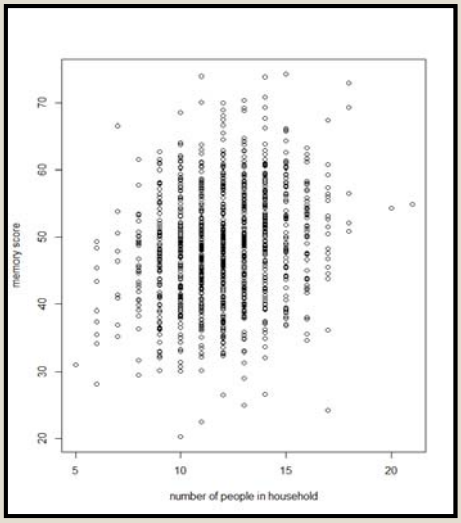
Predictors and covariates

Regression – what is it good for?

- Explore Associations
  - Between outcomes and exposures
  - Between outcomes and exposures adjusting for confounders
  - Hypothesis Testing
- Prediction
  - Generate models to predict an outcome or event
  - Select variables to be included in prediction models
  - Generate "rules" for disease prediction

Dementia and Memory Loss in HIV

- **Factors contributing to memory loss in HIV+ individuals**
- Outcome: Score on memory test
- Exposure/Covariates
  - **Size of household (continuous)**
  - Age (continuous)
  - Sex (binary)
  - Clinic
  - Treatment status (categorical)
  - Years since diagnosis of HIV (continuous)

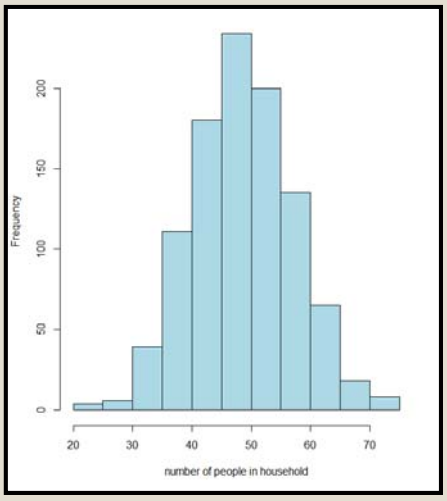


Dementia and Memory Loss in HIV

- **Outcome:**
  - Continuous
  - Normally distributed
- **Simple Linear Regression**

$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$score = \hat{\beta}_0 + \hat{\beta}_1 * size$$



Dementia and Memory Loss in HIV

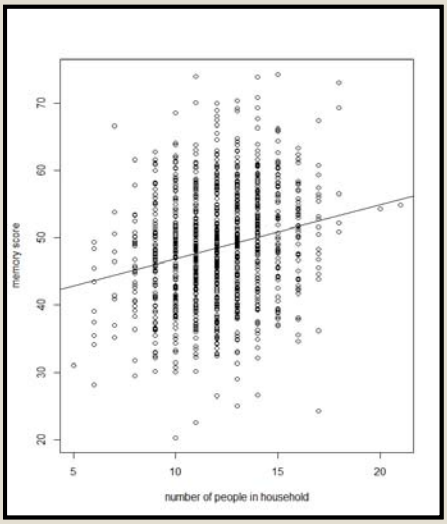
- **Outcome: continuous**
- **Linear Regression**

$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$score = \hat{\beta}_0 + \hat{\beta}_1 * size$$

$$score = 38.7 + 0.81 * size$$

Interpretation?
For each additional person in household, on average the memory score is 0.81 units higher.



Results from linear regression

$\beta_1 = 0.81$, $SE(\beta_1) = 0.11$, $p = 2x10^{-12}$

For each addition person in a household, on aver the score on a memory test is 0.81 units higher.

This association is statistically significant ($p=2x10^{-12}$)

$R^2 = 0.048$

4.8% of the variance in memory scores can be explained by size of household

$$score = 38.7 + 0.81 * size$$

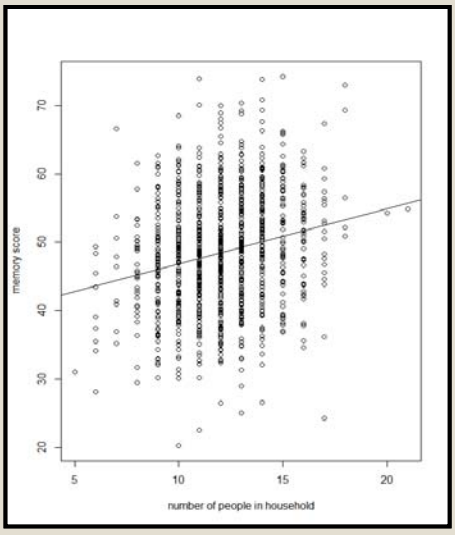
```
> model1 <- lm(full$score ~ full$size_hh)
> summary(model1)

Call:
lm(formula = full$score ~ full$size_hh)

Residuals:
     Min       1Q   Median       3Q      Max
-28.2526  -5.6335   0.1679   5.2436  26.3759

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   38.7303     1.3985  27.694  < 2e-16 ***
full$size_hh   0.8096     0.1137   7.121 2.04e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.276 on 998 degrees of freedom
Multiple R-squared:  0.04836,   Adjusted R-squared:  0.0474
F-statistic: 50.71 on 1 and 998 DF,  p-value: 2.044e-12
```

Linear Regression

- **Factors contributing to memory loss in HIV+ individuals**

$$score = 38.7 + 0.81 * size$$

- Are there other factors leading to memory loss?
- What about possible confounders?
  - Age (continuous)
  - Sex (binary)
  - Clinic (Categorical)
  - Treatment status (categorical)
  - Years since diagnosis of HIV (continuous)



Linear Regression

$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$

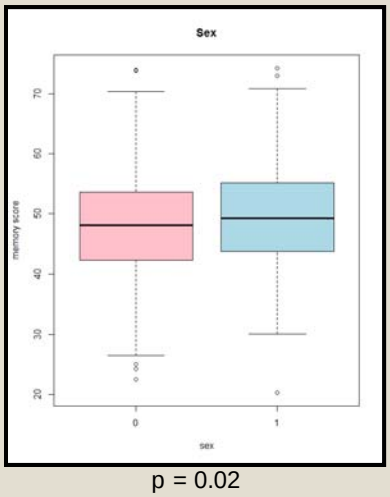
Simple Linear Regression

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$$

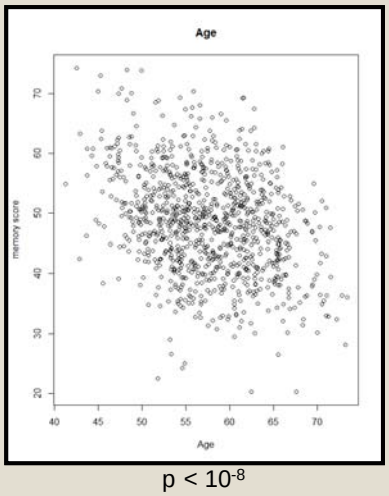
$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k$$

Multiple Linear Regression

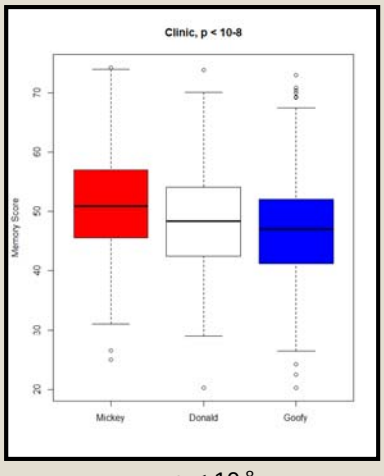
Association with memory score



p = 0.02



$p < 10^{-8}$

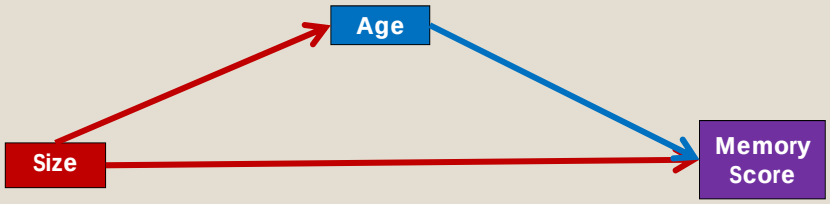


$p < 10^{-8}$

From simple to multiple

- Outcome: Memory Score
- Primary exposure of interest: Size of household
- Covariate: Age

- Does age confound the relationship between Size of household and Memory Score?



Adding Age

	Simple Model		Model with age	
	Beta (SE)	p-value	Beta (SE)	p-value
Size of household	0.81 (0.11)	$2x10^{-12}$	0.41 (0.12)	0.0005
Age			-0.38 (0.05)	$4x10^{-16}$
Adjusted R^2	0.047		0.108	

Questions to ask:

(1) Is size associated with memory score when adjusting for age?

(2) Has the association between household size and memory score changed?

Adding Age

$$score = 38.7 + 0.81 * size$$

$$score = 65.5 + 0.42 * size - 0.38 * age$$

- When controlling for age the association between size of household and score
  - decreased from 0.81 to 0.42 (48%)
  - Remained significant (p=0.0005)
- This suggests that the relationship between household size and score is at least partially mediated through age

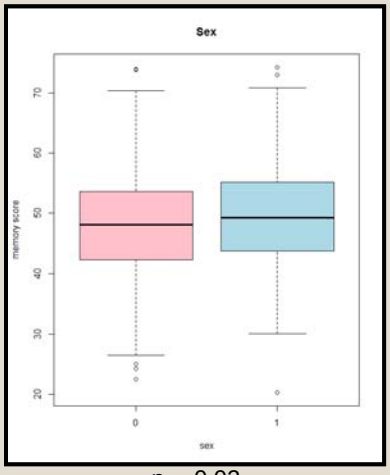
Rule of Thumb: If the β changed by > 10% then consider it a confounder or mediator in the main association

From simple to multiple

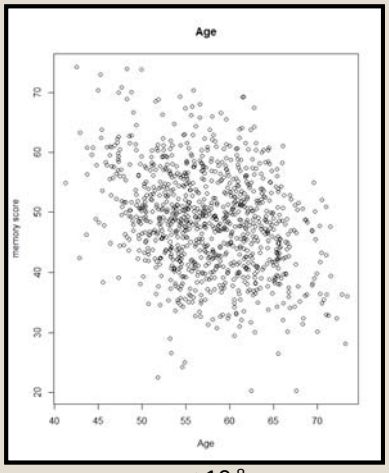
$$score = 65.5 + 0.42 * size - 0.38 * age$$

- "Holding age constant, for each additional person in a household, on average the memory score is 0.42 higher."
- "Among those with the same age, for each additional person in a household, on average the memory score is 0.42 higher."

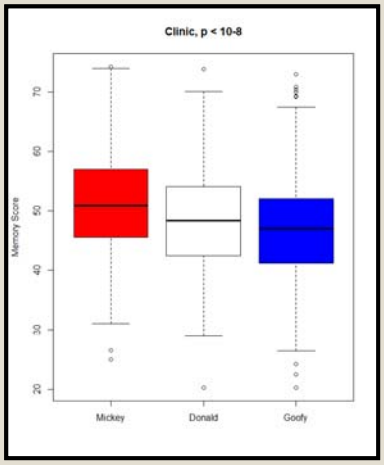
Association with memory score



p = 0.02



$p < 10^{-8}$



$p < 10^{-8}$

Adding Sex

	Simple Model		Model with age		Model with age & sex	
	Beta (SE)	p-value	Beta (SE)	p-value	Beta (SE)	p-value
Size of household	0.81 (0.11)	$2x10^{-12}$	0.41 (0.12)	0.0005	0.42 (0.12)	0.0004
Age			-0.38 (0.05)	$4x10^{-16}$	-0.38 (0.05)	$3x10^{-16}$
Male sex					1.29 (0.52)	0.014
Adjusted R^2	0.047		0.108		0.114	

Questions to ask:

(1) Is size associated with memory score when adjusting for age & sex?

(2) Has the association between household size and memory score changed?

Adding Sex

$$score = 65.5 + 0.42 * size - 0.38 * age$$

$$score = 65.0 + 0.42 * size - 0.38 * age + 1.29 * male$$

- When adding sex to the model with age, the association between size of household and score
  - Did not change (0.42 = 0.42)
  - Household size remained significant (p=0.0004)
  - Sex was significantly associated with memory score (p=0.014)
- Do we leave sex in the model?

Rule of Thumb: If the β changed by > 10% then consider it a confounder or mediator in the main association

Adding Clinic

	Simple Model		Model with age		Model with age & clinic	
	Beta (SE)	p-value	Beta (SE)	p-value	Beta (SE)	p-value
Size of household	0.81 (0.11)	$2x10^{-12}$	0.41 (0.12)	0.0005	0.48 (0.12)	0.00007
Age			-0.38 (0.05)	$4x10^{-16}$	-0.32 (0.05)	$4x10^{-11}$
Clinic 1					1.0	
Clinic 2					-1.29 (0.69)	0.061
Clinic 3					-2.38 (0.64)	0.0002
Adjusted R^2	0.047		0.108		0.118	

Questions to ask:

(1) Is size associated with memory score when adjusting for age & clinic?

(2) Has the association between household size and memory score changed?

Adding Clinic

$$score = 65.5 + 0.42 * size - 0.38 * age$$

$$score = 62.8 + 0.48 * size - 0.32 * age - 1.29 * Clinic2 - 2.38 * Clinic3$$

- When adding clinic to the model with age, the association between size of household and score
  - Changed (0.42 to 0.48)/0.42 = 14%
  - Household size remained significant (p=0.00007)
  - At least one clinic was associated with household size (ps = 0.06, 0.0002)
- Do we leave clinic in the model? Do we leave all 3 clinics in the model?

Rule of Thumb: If the β changed by > 10% then consider it a confounder or mediator in the main association

Additional Confounders

- Duration of HIV and Treatment did **not** change association between household size and memory score

```
>
> model8 <- lm(full$score~full$size_hh + full$age + full$clinicf +  full$HIV_dur
 +full$trt)
> summary(model8)

Call:
lm(formula = full$score ~ full$size_hh + full$age + full$clinicf +
    full$HIV_dur + full$trt)

Residuals:
     Min       1Q   Median       3Q      Max
-26.9915  -5.2785   0.1501   5.0995  22.9114

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       60.01762    3.86568  15.526  < 2e-16 ***
full$size_hh       0.48213    0.12007   4.015 6.38e-05 ***
full$age          -0.32333    0.04865  -6.647 4.94e-11 ***
full$clinicfDonald -1.23042    0.69068  -1.781 0.075143 .
full$clinicfGoofy  -2.31398    0.63806  -3.627 0.000302 ***
full$HIV_dur       0.07506    0.07246   1.036 0.300568
full$trt           1.53500    0.58091   2.642 0.008362 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.939 on 993 degrees of freedom
Multiple R-squared:  0.1287,    Adjusted R-squared:  0.1235
F-statistic: 24.45 on 6 and 993 DF,  p-value: < 2.2e-16
```

Final Model

$$score = 62.8 + 0.48 * size - 0.32 * age - 1.29 * Clinic2 - 2.38 * Clinic3$$

- Larger household size is significantly associated with higher memory score when controlling for age and clinic site.
- For participants with similar age and clinic, for each addition household member, there was a 0.48 higher memory test score.
- Interpretation?

Choosing Covariates

- Use knowledge about your outcome, exposure and research question to help choose covariates to look at. *A priori* knowledge!
- What about automatic selection procedures? (backwards, forwards, stepwise)
  - Based purely on '*the numbers*' (empirical process)
  - Usually only look at significance
  - NOT useful in hypothesis driven testing
- Do not:
  - Test all available variables just because you have them
  - Include highly correlated variables (e.g. height & BMI, right hand strength & left hand strength)

Regression for hypothesis testing

- We did this in Linear Regression, but this process is the same for other regression techniques
- Remember to compare effect size!

- For Logistic, Poisson and Cox Proportional Hazard (Survival Analysis) Regression
  - Compare β estimates

- For ANOVA
  - Compare difference between the group means adjusted for covariates or
  - Model as a linear regression (using dummy variables) and compare β estimates

Regression – what is it good for?

- Explore Associations
  - Between outcomes and exposures
  - Between outcomes and exposures adjusting for confounders
  - Hypothesis Testing

- Prediction
  - Generate models to predict an outcome or event
  - Select variables to be included in prediction models
  - Generate "rules" for disease prediction

Regression for Prediction

For Association

- Typically interested in single (or few) factors that are associated with outcome
- Find other variables that confound that association
- Variables included in the model depending on if they play a role in primary association of interest

For Prediction

- Interested in the 'best set' of variables for a model
- Include all variables that improve **predictive accuracy**
- Less concerned with p-values of specific variables (although p-values are often used to parse down lists of variables)

Predictive Accuracy

- Discrimination: How well the model separates out 'cases' from 'controls'
  - Receiver Operating Characteristic Curve (ROC Curve)
  - Area Under the ROC Curve (AUC or c-statistic)
- Calibration: How well the predicted outcome matches the observed outcome
  - Hosmer-Lemeshow Chi-Square Goodness-of-fit Statistic
- Re-classification: How well a new model to improves on an old model
  - Net Reclassification Index (NRI)
  - Integrated Discrimination Improvement (IDI)
  - Re-classification Index

Dementia and Memory Loss in HIV

- Explore factors that contribute to memory loss in HIV+ individuals
- **Create a prediction rule for onset of dementia**

- Cross-sectional Study of n=1000 HIV+ people
- Collect information on
  - **Dementia diagnosis (binary)**
  - Age (continuous)
  - Sex (binary)
  - Clinic
  - Treatment status (categorical)
  - Years since diagnosis of HIV (continuous)

Outcomes

Predictors and covariates

Dementia and Memory Loss in HIV

- Explore factors that contribute to memory loss in HIV+ individuals
- **Create a prediction rule for onset of dementia (dx_dm)**

- Model: Logistic Regression
- Compare 4 Models
  - Model 1: dx_dm = age
  - Model 2: dx_dm = age + sex
  - Model 3: dx_dm = age + sex + HIV_duration
  - Model 4: dx_dm = age + sex + HIV_duration + HIV treatment + Clinic

$$\text{logit} = \ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_i X_i$$

$$\hat{p} = \frac{e^{\hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_i X_i}}{1 + e^{\hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_i X_i}}$$

Results

	Model 1		Model 2		Model 3		Model 4	
	Beta (SE)	p-value	Beta (SE)	p-value	Beta (SE)	p-value	Beta (SE)	p-value
Age	0.11 (0.02)	10^{-13}	0.11 (0.02)	10^{-13}	0.12 (0.02)	10^{-12}	0.11 (0.02)	10^{-10}
Sex (male)			-0.15 (0.19)	0.41	-0.18 (0.19)	0.36	-0.17 (0.19)	0.357
HIV_dur					0.16 (0.03)	10^{-10}	0.16 (0.03)	10^{-10}
HIV trt							-0.46 (0.20)	0.02
Clinic 2							0.10 (0.28)	0.72
Clinic 3							0.37 (0.25)	0.14
c - stat	0.669		0.671		0.751		0.760	

What to look for:

(1)We are less interested in p-values or if the Betas change

(2) We are looking for the measure of discrimination (c-stat)

c-statistic

$$\hat{p} = \frac{e^{\hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_i X_i}}{1 + e^{\hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_i X_i}}$$

- For every possible case/control pair determine if
  - Concordant, π_c

$$\hat{p}_{case} > \hat{p}_{control}$$

  - Discordant, π_d

$$\hat{p}_{case} < \hat{p}_{control}$$

  - Tie, π_t

$$\hat{p}_{case} = \hat{p}_{control}$$

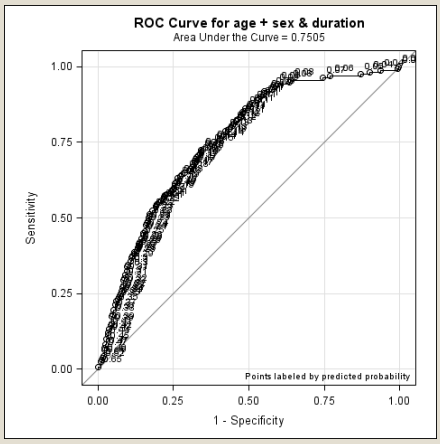
$$c = \frac{\pi_c + \frac{1}{2}\pi_t}{\pi_c + \pi_d + \pi_t}$$

32

ROC Curve

- How to quantify discrimination over all possible thresholds
  - c-statistic (concordance – statistic)
  - Receiver Operating Characteristic (ROC) Curve
  - Area Under Curve (AUC)

- AUC/c-stat ranges from 0.500 to 1.00
  - Null = 0.500
  - Perfect predictive model = 1.00
  - X-Axis = 1-Specificity (False Positive Fraction)
  - Y-Axis = Sensitivity (True Positive Fraction)



Choose a Threshold

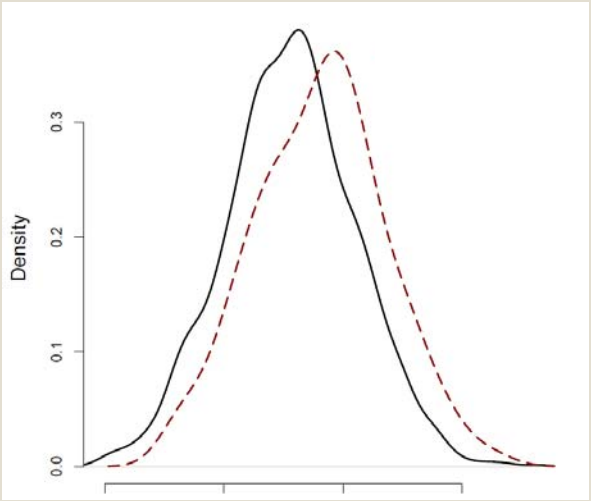
$$\hat{p} = \frac{e^{\hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_i X_i}}{1 + e^{\hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_i X_i}}$$

Remember:
From a logistic regression you can calculate a predicted probability (phat) of event for everyone

If the model predicts the event well, the distribution of phats for events (dotted line) should be to the right of those for non-events (solid line)

A threshold is where everyone with phat > threshold is called “screen pos” and everyone < phat is called “screen neg”

You take your continuous phats, and make a dichotomous screen pos/neg variable



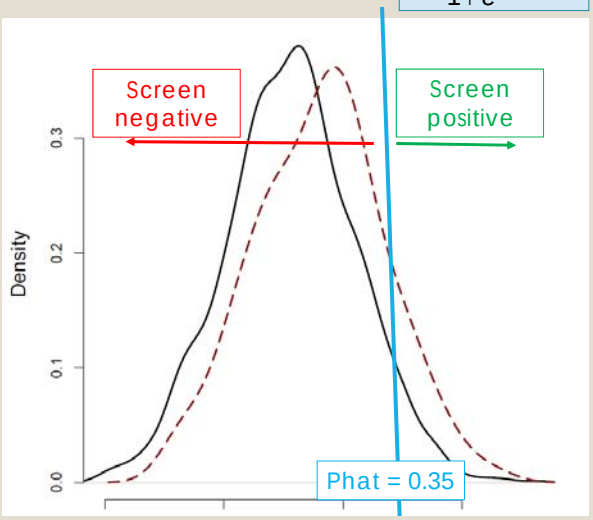
Choose a Threshold

$$\hat{p} = \frac{e^{\hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_i X_i}}{1 + e^{\hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_i X_i}}$$

With a phat threshold = 0.35

If phat > 0.35 => screen positive
If phat < 0.35 => screen negative

Then compare screen pos/neg to actual outcome



Predictive Accuracy

- Sensitivity (TPF) for screening test
  - P(screen = pos | DS=1) = a / (a+c)
- FPF (1-specificity) for screening test
  - P(screen = pos | DS=0) = b / (b+d)

Misclassification Rate = (c+b)/n

		Disease(Y)		
		Yes (1)	No (0)	Total
Screening Test (X)	Pos (1)	a (TP)	b (FP)	a+b
	Neg (0)	c (FN)	d (TN)	c+d
	total	a+c	b+d	n

Predictive Accuracy

- Sensitivity (TPF) for screening test
  - P(screen = pos | DS=1) = a / (a+c)
  - 20 / 160 = **12.5%**
- FPF (1-specificity) for screening test
  - P(screen = pos | DS=0) = b / (b+d)
  - 33 / 840 = **4%**

Misclassification Rate = (c+b)/n
= (33+140)/1000 = 17.3%

		Disease(Y)		
		Yes (1)	No (0)	Total
Screening Test (X)	Pos (1)	20	33	53
	Neg (0)	140	807	947
	total	160	840	1000

37

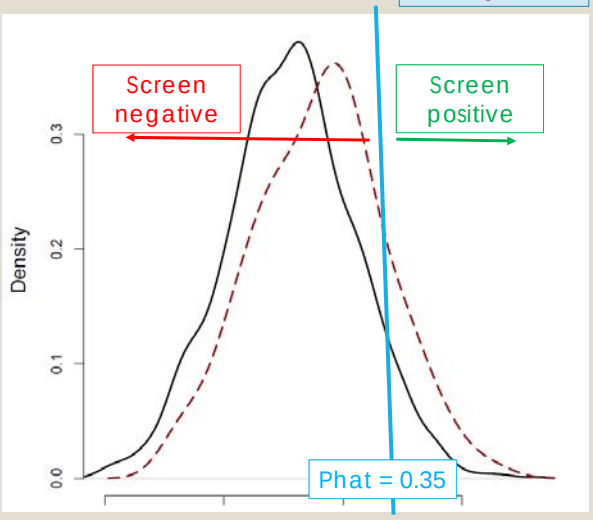
Move the Threshold

$$\hat{p} = \frac{e^{\hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_i X_i}}{1 + e^{\hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_i X_i}}$$

With a phat threshold = 0.35

If phat > 0.35 => screen positive
If phat < 0.35 => screen negative

Then compare screen pos/neg to actual outcome



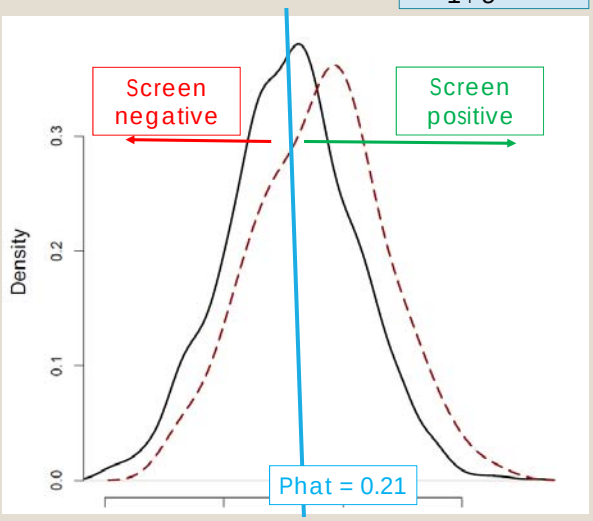
Move the Threshold

$$\hat{p} = \frac{e^{\hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_i X_i}}{1 + e^{\hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_i X_i}}$$

With a phat threshold = 0.21

If phat > 0.21 => screen positive
If phat < 0.21 => screen negative

Then compare screen pos/neg to actual outcome



Predictive Accuracy

- Sensitivity (TPF) for screening test
  - P(screen = pos | DS=1) = a / (a+c)
  - 67 / 160 = **41.9%**
- FPF (1-specificity) for screening test
  - P(screen = pos | DS=0) = b / (b+d)
  - 130 / 840 = **15%**

Misclassification Rate = (c+b)/n
= (93+130)/1000 = **22.3%**

		Disease(Y)		
		Yes (1)	No (0)	Total
Screening Test (X)	Pos (1)	67	130	197
	Neg (0)	93	710	803
	total	160	840	1000

40

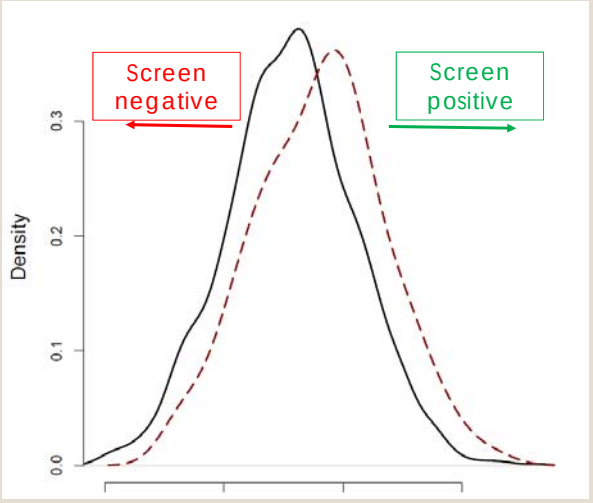
Move the Threshold

$$\hat{p} = \frac{e^{\hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_i X_i}}{1 + e^{\hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_i X_i}}$$

Moving the threshold changes your predictive accuracy measures

Moving threshold higher (right)
- Makes it harder to screen positive
- Decreases sensitivity
- Decreases FPF

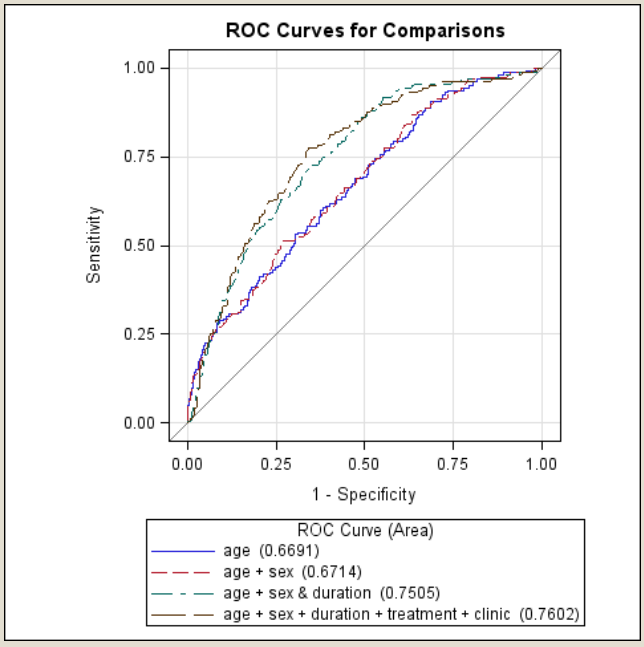
Moving threshold lower (left)
- Makes it easier to screen positive
- Increases Sensitivity
- Increases FPF



ROC Curve

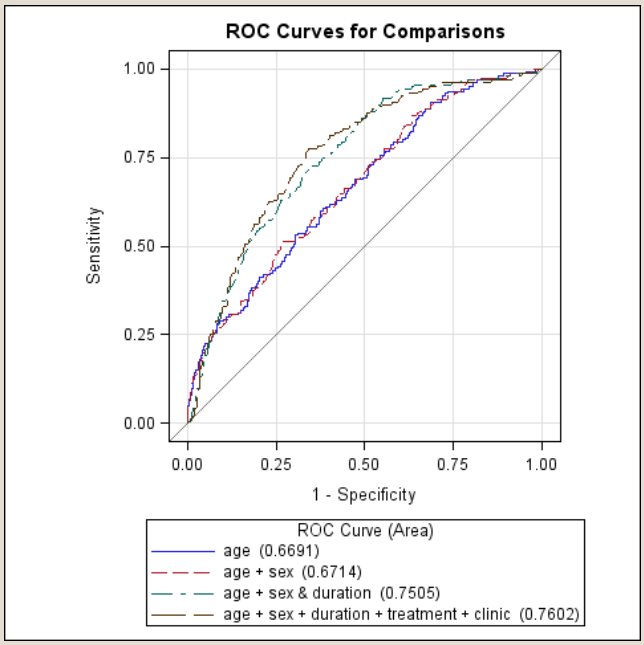
Note:
- "Best" model appears to be the one with age, sex and HIV duration (c=0.75)
- Tiny improvement with addition of trt and clinic (c=0.76), but not statistically different from previous model (p=0.33) - possibly overfitting to data

- Use ROC to help choose variable list but need to test on separate dataset for true assessment of predictive accuracy



ROC Curve

- Good for looking at overall discriminatory ability of models
- Uses continuous predictive probability
- Looks at all possible thresholds for prediction
- By changing thresholds of screen positive or screen negative we change our predictive accuracy



Regression for Prediction: Overfitting

- When we assess predicative accuracy on the same dataset we developed the model in we risk overfitting
- "Overfitting"
  - Imagine taking a mold of your feet and creating the perfect shoe from that mold
  - The shoe will fit great on you, best shoe you ever had
  - How would it fit your neighbor?
- Preventing overfitting
  - Shoe Sizes!
  - Might lose some accuracy, but it is an algorithm that applies to a larger population
- Ideally we have 2 datasets
  - We develop the model on our data (measure all our feet) [training set]
  - Then test it on another dataset (other people's feet) [testing set]

Regression for Prediction: Overfitting

- Ideally we have 2 datasets
  - We develop the model on our data (measure all our feet) [training set]
  - Then test it on another dataset (other people's feet) [testing set]
- Can also do cross-validation
  - Choose 10% of data and set aside
  - "train" the model in the remaining 90%
  - "test" the model in the 10% left out
  - Repeat 10xs and report the distribution of the results (mean, SD)
  - Note: This is 10-fold cross-validation
  - 10 is rather arbitrary – do what you sample size allows, make sure there are enough events/non-events in each set

Regression Summary

For Association

- Typically interested in single (or few) factors that are associated with outcome
- Find other variables that confound that association
- Variables included in the model depending on if they play a role in primary association of interest

For Prediction

- Interested in the 'best set' of variables for a model
- Include all variables that improve **predictive accuracy**
- Less concerned with p-values of specific variables (although p-values are often used to parse down lists of variables)
- Be conscious of overfitting: always test in outside data

QUESTIONS?